8652 NL
ARCHIEF
WA

STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

LEIDING: PROF. DR D. VAN DANTZIG
ADVISEUR VOOR STATISTISCHE CONSULTATIE: PROF. DR J. HEMELRIJK

Rapport S 228 (V 18)

Het begrip nauwkeurigheid

door

Prof. Dr J. Hemelrijk

MATHEMATISCH CENTRUM
Statistische Afdeling

Februari 1958

The Mathematical Centre at Amsterdam, founded the 11th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

1. Inleiding

Eén van de meest hinderlijke, maar onvermijdelijke, eigenaardigheden van bepalingen verricht in een laboratorium is, dat zij bij herhaling vaak verschillende uitkomsten geven. Alleen bij zeer eenvoudige bepalingen (zoals tellingen van een gering aantal objecten) of grove (en daardoor onnauwkeurige) metingen doet dit verschijnsel zich niet of slechts in geringe mate voor. Zodra men echter, in het eerste geval, niet alleen de telling zelf, maar de gehele proef herhaalt - b.v. bij tellingen van bloedlichaampjes niet hetzelfde vakje van een preparaat maar een ander vakje daarvan doortelt - komt het verschijnsel terug. Hetzelfde geschiedt als men, in het tweede geval, ter opvoering van de nauwkeurigheid de meting verfijnt door meer decimalen te gebruiken.

Het is dit verschijnsel, dat het nodig maakt zorgvuldige beschouwingen te wijden aan het begrip nauwkeurigheid. En daarbij maakt men gewoonlijk gebruik van de statistiek, waarmee dan tevens de kansrekening in huis gehaald wordt.

Als eerste benadering kan men zeggen, dat bepalingsmethode A voor een grootheid ξ nauwkeuriger is dan bepalingsmethode B, indien de met A verkregen uitkomsten over het algemeen dichter bij ξ liggen dan die van B. Dit lijkt een triviale uitspraak, maar er schuilt meer in dan men op het eerst gezicht denkt. Het essentiële is nl., dat het nauwkeurigheidsbegrip hier gekoppeld word aan de waarnemingsmethode en niet aan een enkele waarneming.

Het spraakgebruik wil het anders. Men spreekt meestal over de nauwkeurigheid van een waarneming. Deze is op zichzelf gemakkelijk te definiëren: is x een waarneming voor ξ, dan is de absolute afwijking $|x-\xi|$ een theoretisch zeer bevredigende maat voor de onnauwkeurigheid, die echter het nadeel heeft onbekend te zijn, daar ξ dit is (anders is waarneming overbodig). Is dus de (on)nauwkeurigheid van één waarneming gemakkelijk te definiëren, maar practisch van geen nut, omdat hij onbekend is, de nauwkeurigheid van een waarnemingsmethode is moeilijker te definiëren, maar practisch van grote betekenis, daar hij, eenmaal gedefiniëerd, ook experimenteel bepaald kan worden - met grotere of kleinere nauwkeurigheid!

De eigenschappen van een waarnemingsmethode kunnen beschreven worden in de vorm van een kansverdeling. Is ξ de grootheid, die men wil bepalen, terwijl x de uitkomst van een toepassing van de

waarnemingsmethode voorstelt, dan kan $\underline{x}$ allerlei waarden aannemen (waaronder ook ξ). Van deze situatie bouwt men een (theoretisch) model door aan $\underline{x}$ een kansverdeling toe te kennen. Dit wordt aangegeven door het symbool x te onderstrepen: $\underline{x}$.(Het niet onderstreepte symbool x kan dan gebruikt worden om een waarde aan te geven, die $\underline{x}$ aan kan nemen of in een concreet geval aangenomen heeft.)

In de practijk heeft men eigenlijk niet met kansverdelingen te maken, maar met frequentieverdelingen van (altijd) eindige aantallen werkelijk verrichte waarnemingen. Deze kunnen dan gebruikt worden om de kansverdeling van $\underline{x}$ - en daarmede ook ξ en de nog te definiëren nauwkeurigheid - te schatten.

Hoewel het verschil tussen de in de practijk optredende frequentieverdelingen en de theoretische kansverdelingen uitermate belangrijk is, zullen wij op de technische details van de parallele hantering van deze beide niet ingaan. Wij zullen eenvoudig een frequentieverdeling beschouwen als een middel om een kansverdeling te schatten en, als een frequentieverdeling op zeer veel waarnemingen berust, deze in de plaats van de ermee corresponderende kansverdeling stellen. Dat wij de kansverdeling niet geheel kunnen passeren vindt zijn oorzaak in het feit, dat de definitie van nauwkeurigheid daarmee ten nauwste samenhangt.

2. De normale verdeling

De meest bekende kansverdeling is de normale verdeling, die gekarakteriseerd wordt door een kerstklok-vormige kromme, zoals aangegeven in fig. 1.

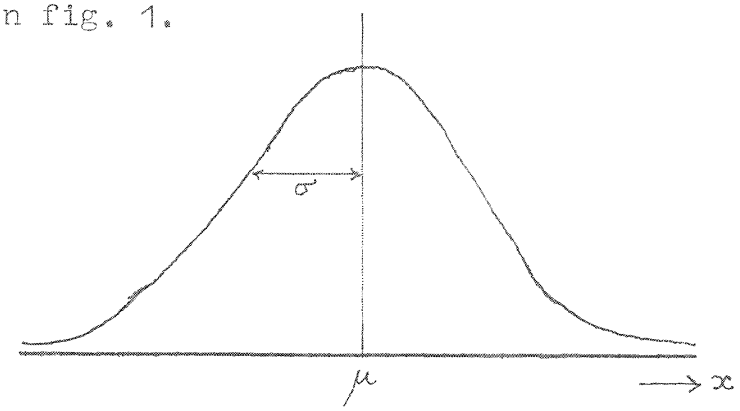


Fig.1. Een normale kansverdeling

Voor ons betekent deze figuur slechts, dat er waarnemingsmethoden zijn, waarvan frequentie-diagrammen, berustend op veel waar-

nemingen, ongeveer de geschetste vorm aannemen. Dit is, gelukkig, het geval bij een uitgebreide categorie van waarnemingsmethoden, zoals titraties, wegingen, lengtemetingen, enz., ieder vaak herhaald aan hetzelfde object.

Een dergelijk kromme wordt volledig gekarakteriseerd door twee grootheden ("parameters"). De eerste daarvan is μ, het symmetriepunt, centrum, zwaartepunt, gemiddelde van de verdeling (of: de mathematische verwachting van x). Deze parameter geeft grofweg de ligging aan. De tweede parameter is σ, de afstand van μ tot de plaats, waar de kromme een buigpunt vertoont. Deze parameter wordt de standaardafwijking of de spreiding genoemd en hij karakteriseert de nauwkeurigheid van de waarnemingsmethode, beschouwd als waarnemingsmethode voor μ. Een grote σ betekent, dat grote afwijkingen van μ vaak voorkomen, een kleine, dat dit zelden gebeurt. Voor normale verdelingen geldt verder de in de statistiek zeer bekende stelling, dat afwijkingen van μ groter dan 2σ slechts bij ongeveer 5% van de toepassingen der beschouwde methode zullen optreden [1]. Een aanschouwelijke en practisch zeer goed bruikbare maat voor de (on)nauwkeurigheid van een waarnemingsmethode is dus σ zelf of 4σ, de lengte van het interval, waarin ongeveer 95% der waarnemingen, bij herhaalde toepassing, zullen vallen.

Is σ bekend, en is één waarneming x verkregen, dan kan men dus, met onbetrouwbaarheid 0,05, beweren dat voor μ geldt:

$$(1) \qquad x-2\sigma \leq \mu \leq x+2\sigma.$$

Een bekend en vaak toegepast middel om de nauwkeurigheid te vergroten is: meer dan één waarneming te verrichten. Wordt de meting n maal in zijn geheel herhaald, terwijl daarbij de uitkomsten $x_1, x_2, \ldots, x_n$ verkregen worden, dan geldt i.p.v. (1), met $\bar{x} = \frac{1}{n}\Sigma x_i$:

$$(2) \qquad \bar{x} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2\frac{\sigma}{\sqrt{n}}$$

met dezelfde betrouwbaarheid van 95%.

(Daarbij moet het voorbehoud gemaakt worden, dat de waarnemingen onafhankelijk zijn, d.w.z. dat zij elkaars uitkomsten niet beïnvloeden; dit wordt bereikt door telkens dezelfde handelingen zoveel mogelijk op dezelfde wijze uit te voeren, zonder daarin

---

1) Preciezer: $1{,}96\sigma$.

enige verandering aan te brengen naar aanleiding van reeds verkregen uitkomsten. Zo is het b.v. onjuist uitkomsten, die van vroegere nogal sterk afwijken om die reden weg te laten bij de berekening van $\bar{x}$.)

Daar meestal ook σ onbekend is, wordt deze parameter eveneens uit de verrichte waarnemingen geschat, volgens de formule:

$$(3) \qquad s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \ .$$

Is het aantal waarnemingen gering, dan dient tevens in (2) de factor 2 door een andere (gewoonlijk met t aangegevene) vervangen te worden, die in een tabel van de z.g. STUDENT-verdeling kan worden opgezocht. Formule (2) wordt dan

$$(4) \qquad \bar{x} - t \frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + t \frac{s}{\sqrt{n}} \ .$$

Deze formule geldt voor iedere n. Voor kleine n is $t > 2$, voor grote n verschilt t niet noemenswaard van 2.

Een interval als door (1), (2) of (4) gegeven, wordt een <u>betrouwbaarheidsinterval</u> genoemd en wel zijn hier alleen betrouwbaarheidsintervallen met onbetrouwbaarheid 0,05 genoemd. Men kan echter voor deze onbetrouwbaarheid iedere gewenste waarde nemen, waarbij verkleining plaats vindt ten koste van vergroting van het interval en v.v.

Zijn uit vroegere toepassingen der methode meer gegevens over σ te halen, dan kan men veelal σ als voldoende bekend beschouwen om (2) toe te passen.

Op deze wijze kan de nauwkeurigheid van $\bar{x}$ als schatting van μ op bevredigende wijze beschreven worden. De beschrijving heeft weer betrekking op de <u>methode</u> (het nemen van het gemiddelde van een aantal onafhankelijke waarnemingen), niet op één enkele uitkomst voor $\bar{x}$. In ieder concreet geval is de verkregen uitspraak òf goed, òf fout. Bij herhaalde toepassing geeft de onbetrouwbaarheid aan, hoe vaak men een fout resultaat verkrijgt, maar niet, wèlke resultaten fout zijn.

3. Systematische fouten

Hiermede is het doel echter nog geenszins bereikt. Immers wij zochten naar de nauwkeurigheid van een waarnemingsmethode voor een grootheid ξ en wij hebben er nu één voor μ. Is

(5) $$\xi = \mu ,$$

dan hebben wij ons doel dus bereikt, maar als

(6) $$\xi \neq \mu$$

is, dan is dit niet het geval. Dit aspect van de zaak wordt gewoonlijk afzonderlijk beschouwd. Is (5) vervuld, dan wordt de waarnemingsmethode zuiver genoemd, is (6) daarentegen juist, dan onzuiver; in het laatste geval spreekt men ook van het optreden van een systematische fout (en wel van de grootte $\mu - \xi$).

Een dergelijke systematische fout kan b.v. veroorzaakt worden door ontregeling van de meetapparatuur, door verkeerde instelling van een aflees-schaal, door het optreden van een parallaxe, door een verkeerde concentratie van een oplossing, of, in het algemeen, door een telkens herhaalde verkeerde handeling bij de waarneming. Indien dit het geval is - en het is vaak moeilijk dit geheel te vermijden - dan is de gebruikte waarnemingsmethode (inclusief deze systematische fout!) eigenlijk in meerdere of mindere mate ongeschikt voor waarneming van ξ. Immers μ wordt waargenomen en μ is $\neq \xi$ Zijn μ en ξ bij het verrichten van een serie metingen constant (d.w.z. treedt er geen systematische verandering in de meetmethode of in de te meten grootheid op tijdens het verrichten van de reeks waarnemingen), dan zal, als n groot genoeg wordt, ξ zelfs vrijwel zeker buiten de intervallen (2) en (4) komen te liggen. Hieruit blijkt, dat deze intervallen dan niet geschikt zijn voor de schatting van ξ en dat het zinloos is een groot aantal waarnemingen te verrichten met dezelfde apparatuur (waarbij inbegrepen is de waarnemer, de bediener van de apparatuur), indien men zich er niet zorgvuldig van overtuigd heeft, dat aan (5) is voldaan. Alleen zolang een eventueel mogelijk verschil tussen μ en ξ klein is in vergelijking met de lengte ($4 \frac{\sigma}{\sqrt{n}}$ resp. $2t \frac{s}{\sqrt{n}}$) van het betrouwbaarheidsinterval, is het zinvol dit interval ook op ξ te betrekken.

De vraag doet zich nu voor, hoe men na kan gaan of ξ gelijk is aan μ of niet en in het laatste geval, hoe groot het verschil is.

Dit geschiedt door ijking van de meetmethode: toepassing ervan op één of meer objecten met bekende ξ . De uitkomsten, bij herhaalde toepassing moeten dan symmetrisch om ξ gespreid liggen en dit kan met behulp van de toets van STUDENT voor één steekproef onderzocht worden. Als $\mu \neq \xi$ blijkt te zijn kan door nauwkeurig onderzoek de grootte van het verschil bepaald worden, zodat een correctie aangebracht kan worden waardoor de waarnemingsmethode van een onzuivere in een zuivere overgaat. Dit kan echter alleen als het verschil constant is, zodat de correctie ook voor later waarnemingen geldig blijft.

Veelal is echter de systematische fout van verschillende factoren afhankelijk: verschillend voor verschillende waarnemers, veranderlijk in de tijd, afhankelijk van het gebruikte exemplaar van het instrument, enz. In dat geval kan men μ zelf als een stochastische grootheid beschouwen, waarvan dan veelal verondersteld wordt, dat deze - nu beschouwd voor een verzameling (populatie) van gelijksoortige waarnemingsmethoden - opnieuw een normale verdeling bezit.

Deze situatie doet zich b.v. voor indien men een bloedmonster in delen verdeelt en deze ter analyse naar verschillende laboratoria opstuurt ter bepaling van, laten we zeggen, het alcoholgehalte [2]. Aannemende - hetgeen ook nog niet eens waar behoeft te zijn - dat ξ (het gezochte gehalte) in de verschillende porties, waarin het monster gesplitst is, op het tijdstip van analyse dezelfde waarde bezit, is voor de verschillende laboratoria de waarde van μ toch verschillend. Is de standaardafwijking van deze μ's gelijk aan σ_μ en die van $\underline{x}$ (om de bijbehorende μ) voor alle laboratoria gelijk aan σ , dan is de standaardafwijking van het gemiddelde van alle verrichte metingen gelijk aan

$$(6) \qquad \sqrt{\frac{\sigma_\mu^2}{k} + \frac{\sigma^2}{n\,k}}$$

waarin k het aantal laboratoria voorstelt, terwijl in ieder daarvan n bepalingen zijn verricht. In overeenstelling met wat reeds eerder gezegd is blijkt uit deze formule, dat het zinloos is n groot te ne-

---

2) Vgl. Chr.L. RÜMKE en A.R. Bloemena, Statistische beschouwingen over de bloedproef II, Verkeersrecht 5 (1957), 233-236.

men zonder dit ook voor k te doen, tenzij σ_μ^2 veel kleiner is dan σ^2.

Voor een gegeven totaal aantal waarnemingen N=nk wordt de uiteindelijke standaardafwijking (6) zo klein mogelijk als men k zo groot mogelijk maakt, dus aan àlle daarvoor beschikbare laboratoria een deel van het monster stuurt. Dit geldt echter alleen als σ voor alle laboratoria gelijk is. Is over de σ der laboratoria weinig bekend, dan zal men in ieder geval van elk laboratorium minstens 2 onafhankelijke bepalingen moeten verlangen. Bij dit alles kunnen zich nog vele complicaties voordoen, waarvan wij er slechts één noemen. Deze is, dat gewoonlijk niet veel laboratoria beschikbaar zijn en dat dientengevolge k in de regel zo klein zal zijn, dat de verschillen tussen μ en ξ voor deze laboratoria bij optelling zelfs niet bij benadering tegen elkaar weg behoeven te vallen. Er blijft dan ook nu nog een systematische fout over. In laatste instantie kan die alleen door zorgvuldige ijking van de apparatuur worden bestreden en niet door het nemen van een gemiddelde over de uitkomsten van verschillende laboratoria.

4. Tellingen

Wij komen nu terug op de in het begin reeds genoemde tellingen en wel beschouwen wij tellingen van bloedlichaampjes, die de één of andere afwijking vertonen; met behulp van hun aantal (per mm^3 bloed) bepaalt men b.v. of de proefpersoon aan een bepaalde ziekte lijdt of niet. Deze telling geschiedt onder het microscoop in een telkamer, na verdunning van het bloed, daar er anders teveel bloedlichaampjes in het gezichtsveld vallen.

Hier doen zich dezelfde kwesties voor als boven. Bij herhaling van de telling aan een tweede druppel van hetzelfde bloedmonster of aan een tweede bloedmonster van dezelfde proefpersoon wordt in de regel een andere uitkomst verkregen dan eerst. Systematische fouten kunnen optreden door factoren als: onzorgvuldig tellen (geeft een te laag aantal), onnauwkeurig verdunnen (kan zowel een te hoge als een te lage uitkomst geven), verkeerde inhoud van de telkamer, enz. Tegen deze fouten moet men zich weer door ijking te weer stellen.

Daarnaast heeft men ook nu weer te maken met toevallige schommelingen, die nu echter - in het bijzonder als de tellingen in

kleine aantallen resulteren - niet meer tot een normale verdeling leiden, maar tot een verdeling van POISSON. De aard van deze verdeling wordt door één parameter bepaald, die we weer met μ aangeven. Zijn een aantal tellingen met dezelfde apparatuur verricht aan één bloedmonster, dan is het gemiddelde van de verkregen uitkomsten weer een schatting van μ, die nauwkeuriger is dan het resultaat van één afzonderlijke telling.

Ook nu kan de nauwkeurigheid van de methode - afgezien van eventuele systematische fouten - weer met behulp van een betrouwbaarheidsinterval worden uitgedrukt. De bepaling daarvan geschiedt echter niet met een eenvoudige formule, zoals (2) of (4), doch op grond van tabellen of nomogrammen van POISSON-verdelingen.

Indien μ het gemiddelde aantal bloedlichaampjes met de beschouwde afwijking per mm^3 bloed van de proefpersoon voorstelt (op het moment van de bloedafname), terwijl een volume van δ mm^3 bloed onder het microscoop wordt doorgeteld, dan is de parameter van de POISSON-verdeling, waaraan de tel-uitkomst ontleend wordt, gelijk aan $\delta\mu$. In tabel 1 wordt, voor verschillende waarden van de tel-uitkomst (x), het bijbehorende betrouwbaarheidsinterval voor $\delta\mu$ gegeven, met wederom betrouwbaarheid 0,95.

Tabel 1

Betrouwbaarheidsinterval (0,95)
voor de parameter $\delta\mu$ van een Poisson-verdeling

geteld aantal in volume δ	ondergrens voor $\delta\mu$	bovengrens voor $\delta\mu$
3	0,6	8,8
4	1,0	10,3
5	1,6	11,7
6	2,2	13,1
7	2,8	14,5
10	4,7	18,5
15	8,4	24,8
20	12,2	30,8

De grote onzekerheid, waarmede dergelijk tellingen behept zijn, blijkt duidelijk uit deze tabel. De overeenkomstige betrouwbaar-

heidsintervallen voor μ zelf worden verkregen door de grenzen voor $\delta\mu$ door δ te delen, waarbij dan afgezien is van eventuele onnauwkeurigheden bij de verdunning. Verricht men meerdere tellingen, b.v. in verschillende telvakjes, dan neemt men deze gewoon tezamen en telt de bijbehorende δ's eveneens op. Men komt dan bij een hoger aantal uit en het betrouwbaarheidsinterval voor μ wordt korter. Dit is gemakkelijk aan de hand van numerieke voorbeelden na te gaan.

Tot zover loopt alles dus geheel parallel aan het bij de normale verdeling behandelde. Wenst men nu echter nog een tweede foutenbron, b.v. de verdunning in de beschouwingen te betrekken, dan wordt het probleem veel ingewikkelder. Eenvoudige formules, zoals (6), zijn nu niet van toepassing en de oplossing van dit probleem gaat buiten het bestek van deze voordracht.

5. De variatiecoëfficiënt

De nauwkeurigheidsmaat, die wij tot nu toe besproken hebben: de lengte van een betrouwbaarheidsinterval met betrouwbaarheidscoëfficiënt 0,95, is een absolute maat. Daarnaast wordt de variatiecoëfficiënt vaak als een relatieve maat voor de nauwkeurigheid gebruikt. Deze wordt gedefinieerd als

$$v = \frac{\sigma}{\mu} \tag{7}$$

en hij heeft weer alleen betrekking op toevallige afwijkingen ten opzichte van μ, niet op systematische fouten.

Het belang van deze maat zullen wij aan een laatste voorbeeld duidelijk maken.

Indien men een breuk

$$q = \frac{x}{y} \tag{8}$$

beschouwt, waarna teller en noemer beide met onnauwkeurigheden behept zijn (d.w.z. stochastische grootheden zijn), dan hangt de onnauwkeurigheid van q sterk tezamen met de relatieve onnauwkeurigheid van teller en noemer. Zonder hiervan een bewijs te leveren willen wij dit aan een numeriek voorbeeld toelichten.

Wij beschouwen daartoe - voor de eenvoud - grootheden, die slechts twee waarden aan kunnen nemen, met gelijke kansen.

Is x_1 een dergelijke grootheid, die de waarden 2 en 5 aan kan

nemen, dan kunnen wij het quotiënt van deze beide (2/5) beschouwen als een maat voor de relatieve nauwkeurigheid. Dit is niet de variatiecoëfficiënt volgens (8), maar een daarmee op eenvoudige wijze samenhangende grootheid, die bij dit voorbeeld de voorkeur verdient vanwege zijn aanschouwelijkheid.

Kan y_1 de waarden 1 en 2 aannemen (relatieve nauwkeurigheid $\frac{1}{2}$), dan zijn de uiterste waarden, die x_1/y_1 aan kan nemen: 2/2=1 en 5/1 =5; het quotiënt hiervan kunnen wij - de tussenliggende mogelijkheden even buiten beschouwing latende - weer beschouwen als een maat voor de relatieve nauwkeurigheid van de breuk x_1/y_1. Deze is dan $\frac{1}{5}$.

Laat nu x_2 de waarden 2γ en 5γ aannemen in plaats van 2 en 5, zodat de relatieve nauwkeurigheid van $\underline{x}_2$ dezelfde is als die van $\underline{x}_1$, dan zijn de uiterste waarden van x_2/y_1: γ en 5γ weer met quotiënt 1/5. Kan y_2 de waarden ζ en 2ζ aannemen, dan x_1/y_2 als uitersten $1/\zeta$ en $5/\zeta$ en x_2/y_2: y/ζ en $5y/\zeta$. Het quotiënt blijft steeds $\frac{1}{5}$.

Dit doet vermoeden - hetgeen, indien teller en noemer onafhankelijk zijn, ook in het algemeen bij benadering juist is - dat de relatieve nauwkeurigheid van een breuk niet verandert indien die van teller en noemer niet veranderen. Voor de variatiecoëfficiënt geldt dit - bij benadering - ook, zolang die van de noemer niet te groot is. De zaak ligt dan echter wat ingewikkelder dan hier beschreven is.

Brengen wij echter veranderingen aan in de door teller en noemer aan te nemen waarden, waarbij niet de relatieve, maar de absolute nauwkeurigheid onveranderd blijft, dan ziet het er heel anders uit. De absolute (on)nauwkeurigheid kunnen wij meten aan het verschil der twee door x_1 (resp. y_1) aan te nemen waarden: 5-2=3 (resp. 2-1=1). En voor x_1/y_1: 5-1=4. Laat nu b.v. y_2 de waarden 0,01 en 1.01 aannemen (weer met het verschil 1), dan zijn de uiterste waarden van x_1/y_2: 2/1,01≈2 en 5/0,01=500. Zowel de relatieve als de absolute onnauwkeurigheid zijn nu enorm toegenomen. Analoge beschouwingen gelden voor de teller.

Zodra men met breuken te maken heeft is dan ook de relatieve nauwkeurigheid - en in het bijzonder de variatie als maat daarvan - van groot belang.

Ter illustratie bespreken wij [3] een passage uit het proefschrift van K.H. BRANDT: "Over de plaats van vorming der urobilinogenen in het menselijk organisme", Utrecht 1957.

De passage betreft het verschil tussen de z.g. B-gal, die wordt uitgescheiden na de galblaas te zijn gepasseerd, en de C-gal welke rechtstreeks uit de lever komt.

Doordat de gal in de galblaas door resorptie van water wordt ingedikt is bij eenzelfde persoon de bilirubineconcentratie in de B-gal aanmerkelijk hoger dan in de C-gal. Bij 21 patiënten [4] van BRANDT varieerde de verhouding van beide concentraties van ruim 1,5 tot 50 [5]. Ook voor de urobilinogenen vond BRANDT bij al deze personen, op één uitzondering na, de hoogste concentratie in de B-gal.

Van de waarnemingen, bij deze 21 personen verricht (l.c., p. 116-117 [6]), zijn enkele gegevens in tabel 2 overgenomen. Daarbij zijn concentraties, zoals gebruikelijk, door vierkante haken aangegeven.

---

3) De volgende beschouwingen zijn ontwikkeld in samenwerking met Prof.Dr C.G.G. VAN HERK, medewerker van de Statistische Afdeling van het Mathematisch Centrum, Amsterdam.

4) Bijna allen hadden een onbelaste anamnese wat betreft lever- en galwegen; één patiënt had een steen, één had een hepatitis infectiosa doorgemaakt, maar tijdens het onderzoek een goede leverfunctie.

5) Misschien zijn deze zeer uiteenlopende getallen ten dele toe te schrijven aan een meer of minder sterke vermenging van gal met vocht in het duodenum. Zolang dit vocht zelf geen galkleurstoffen bevat is een dergelijke verdunning voor de verdere discussie van geen belang..

6) Om ons onbekende redenen zijn de gegevens over patiënt 21 hier niet vermeld. Deze zijn ons door de schrijver verstrekt en ook in deze beschouwingen opgenomen. Weglating zou echter geen essentiële wijziging tot gevolg hebben. De tabel bevat enkele onbetekende rekenfouten.

Tabel 2

Concentraties van urobilinogenen (in mg/100 cc) en bilirubine (in E/100 cc) in B- en C-gal

nr. patiënt	$[ur]_B$	$[ur]_C$	$[bil]_B$	$[bil]_C$
1	0,37	0,06	96,96	15,70
2	0,108	0,005	99,26	2,82
9	0,312	0,208	98,7	8,7
12	1,38	0,57	46	9,74
19	0,01	0,005	20,6	13,1

Van belang is verder dat geen bilirubine in de galblaas wordt gevormd of afgescheiden. Wel kan deze stof daar verdwijnen, althans onder pathologische omstandigheden, maar voor ons betoog speelt dit (gelukkig) geen rol.

Voor iedere patiënt beschouwt BRANDT nu de beide quotiënten:

(9) $$q_{bil} = [bil]_B / [bil]_C \ , \ q_{ur} = [ur]_B / [ur]_C \ .$$

Indien in de galblaas alleen indikking plaats vindt is er geen reden waarom deze quotiënten onderling zouden verschillen, behalve dan door onvermijdelijke onnauwkeurigheden bij de chemische analyse [7]. Zou echter blijken dat q_{ur} systematisch kleiner is dan q_{bil}, dan zou dit wijzen op resorptie of afbraak van urobilinogenen in de

---

7) Met een mogelijke storende invloed van het leverrhythme op de samenstelling van de gal is geen rekening gehouden. Hierover hebben wij nl. geen gegevens kunnen verkrijgen. JORES (Tabulae biologicae 14 ('37) p. 95-98) geeft een dagkromme voor het bilirubinegehalte van het bloed (maximum om ca. 0^h, minimum ca. 4^h) en voor de urobilinogeenuitscheiding in de urine (maximum ca. 12^h, minimum ca. 4^h), maar spreidingen worden daarbij niet vermeld. Bij het gebruikelijke onderzoek worden de B- en C-gal op verschillende delen van een etmaal gesecerneerd. Voor de statistische aspecten van deze discussie is dit overigens van geen belang.

galblaas. Omgekeerd zou door uitscheiding of aanmaak van urobilinogenen (al dan niet bilirubine) in de galblaas, q_{ur} systematisch groter worden dan q_{bil} [8].

Statistisch bezien komt het er dus op neer of van de beide quotiënten het ene bij een voldoend aantal patiënten voldoende kleiner is dan het andere. Dit houdt nl. een aanwijzing in voor een systematisch verschil, en daarmee voor ten minste één der verschijnselen: resorptie, aanmaak etc. van hetzij urobilinogenen, hetzij bilirubine in de galblaas.

In fig. 2 zijn de beide concentratiequotiënten voor deze 21 patiënten grafisch tegen elkaar uitgezet (zie diss. p. 131).

Ligt een punt in deze figuur boven de bissectrice door de oorsprong, dan is $q_{ur} > q_{bil}$, terwijl voor punten eronder het omgekeerde geldt. Geen der punten ligt precies op deze lijn, wat trouwens ook niet kon worden verwacht.

Van de 21 punten liggen er nu 6 boven en 15 onder de bissectrice. Evenals bij een eerder genoemd voorbeeld verkeren wij daarom in een situatie waarin geen voldoende aanwijzingen voorhanden zijn voor een systematisch verschil, hoewel fig. 2 toch wel een vermoeden in die richting kan doen ontstaan.

Nu wijken een aantal punten, zowel boven als onder de bissectrice, ver van deze lijn af. Dit doet vermoeden dat de nauwkeurigheid der quotiënten gering is. Immers als in de galblaas alleen indikking plaats vindt, zijn alle afwijkingen van de bissectrice aan onnauwkeurigheden van de analyse te wijten. Maar ook als dit niet het geval is en er b.v. ook afbraak van urobilinogenen plaats vindt, zijn deze punten met een redelijk uniforme galblaasfunctie van de onderzochte personen onverenigbaar. De verhouding q_{ur}/q_{bil} varieert van 0,132 tot 5,74, zodat de grootste waarde hiervan ruim 43 maal de kleinste bedraagt. Voor personen met ogenschijnlijk normale lever en galwegen, - zij het ook met mogelijke andere afwijkingen - , lopen deze getallen toch wel erg ver uiteen om zonder meer aan functieverschillen te worden toegeschreven. Er is trouwens een veel meer voor de hand liggende verklaring.

Over de nauwkeurigheid der bepalingen waren wij niet ingelicht,

---

8) Dit zou ook het geval zijn bij afbraak of selectieve resorptie van bilirubine alleen.

vermoedelijk was deze onbekend. Deze onbekendheid, - een lacune in de opzet van het onderzoek -, had tot gevolg dat de statistiek niet als bewijsmiddel kon worden gehanteerd, zodat het waarnemingsmateriaal geen duidelijke conclusie toeliet.

Maar het is niet onaannemelijk, - zoals nader zal blijken -, dat dit anders zou zijn geweest indien ook de nauwkeurigheid van de analyse behoorlijk was onderzocht, o.a. door proeven te nemen met afgewogen hoeveelheden urobilinogeen en door het verrichten van een voldoend aantal duplowaarnemingen. Bij ontbreken hiervan kan de statistiek altijd nog fungeren als detectiemiddel.

Redelijke speculaties over de analysenauwkeurigheid kunnen nl. het effect, dat in fig. 2 in zwakke mate aanwezig schijnt te zijn, duidelijker naar voren brengen. Als regel is immers de relatieve nauwkeurigheid van een quantitatieve bepaling (bij eenzelfde methodiek) geringer, wanneer het kleinere hoeveelheden van de onderzochte stof betreft. De relatieve verliezen zijn dan i.h.a. groter, de meeste aflezingen naar verhouding onnauwkeuriger.

Hieruit volgt, dat ook de verhouding van twee hoeveelheden van een zelfde stof des te onnauwkeuriger wordt bepaald, naarmate deze hoeveelheden kleiner zijn. Omgekeerd zal van twee ongeveer gelijke, aldus verkregen quotiënten datgene met de grootste teller en noemer het betrouwbaarst zijn.

Bovendien zijn de door verliezen veroorzaakte fouten éénzijdig. Bij de zo juist beschouwde quotiënten zullen daarom, in het algemeen gesproken, zowel de teller als de noemer te klein uitvallen. Is één van deze grootheden aanmerkelijk kleiner dan de andere, dan zal de relatieve invloed van de verliezen op het kleinste getal het sterkst zijn. Is dit getal de noemer, dan zal voor het quotiënt een te grote, en soms zelfs een exorbitant grote waarde worden gevonden.

Wij vermoeden dat een en ander zich bij de boven beschouwde grootheden voordoet, speciaal bij de quotiënten q_{ur}. Immers de bilirubineconcentratie is in de gal groter dan die van het urobilinogeen, en de voor q_{bil} verrichte aflezingen waren blijkbaar steeds in ten minste 2 cijfers mogelijk; vermoedelijk is daarom q_{bil} nauwkeuriger bepaald dan q_{ur}.

Daarentegen is in het bijzonder de noemer $[ur]_C$ van q_{ur} soms zeer klein (d.w.z. in niet meer dan één cijfer aangegeven), en vermoedelijk dus zeer onnauwkeurig. Dat de drie kleinste waarden van

$[ur]_c$ in de (volledige) tabel alle gelijk zijn aan 0,005 ondersteunt dit vermoeden in niet geringe mate. Wellicht liet de chemische analyse een verdergaande specificatie van het begrip "zeer klein" niet toe.

Het ligt voor de hand dat ecarteren van de minst goede waarnemingen een onderzoek als het nu besprokene ten goede kan komen. De onnauwkeurige waarnemingen kunnen een eventueel aanwezig effect gemakkelijk verdoezelen.

De meest voor de hand liggende wijze van ecarteren bestaat in het weglaten van die gevallen waarbij $[ur]_c$ klein en q_{ur} dus onbetrouwbaar is. Wel is het gewenst dat daarbij genoeg punten uit fig. 2 overblijven om over de structuur van de resterende puntenwolk een behoorlijke indruk te kunnen krijgen. Nu is bij 13 van de 21 beschouwde patiënten $[ur]_c \leqslant 0,104$; bij de 8 overigen varieerde deze grootheid van 0,208 tot 1,31 [9]. De met deze laatsten corresponderende betrouwbaardere punten zijn in fig. 3 weergegeven.

Deze liggen nu alle duidelijk beneden de bissectrice, hetgeen een sterke aanwijzing voor een systematisch verschil tussen q_{bil} en q_{ur} zou inhouden, als bovenstaande procedure verantwoord kon worden geacht. Was een uitkomst van de aard van fig. 3 verkregen na ecartering op grond van nauwkeurigheidsbepalingen, dan was een ondubbelzinnig resultaat bereikt.

Bij de gevolgde methode verdwijnen evenwel juist die punten, waarvoor $[ur]_c$ klein, dus q_{ur} i.h.a. groot is, d.w.z. overwegend hooggelegen punten in fig. 2. En inderdaad zijn ook de hooggelegen punten onder de bissectrice verdwenen. Toch was het niet vanzelfsprekend dat de overgebleven punten alle onder de bissectrice zouden liggen; q_{ur} kan nl. óók groot zijn door een grote teller (waar over later). Maar het trekken van een conclusie uit het verkregen resultaat alleen is zeker niet verantwoord.

Men kan echter trachten op grond van andere criteria te ecarteren, en zien of men aldus aanwijzingen verkrijgt die in dezelfde richting gaan. Dit blijkt inderdaad het geval. Daarbij valt te bedenken dat iedere methode van ecarteren willekeurig is (zoals de grens voor $[ur]_c$, die aan fig.3 ten grondslag lag, vrij arbitrair

---

9) De lacune tussen 0,104 en 0,208 in de waarden van $[ur]_c$ verschafte een ongedwongen criterium voor het ecarteren.

was). Er zijn immers geen voldoende gegevens over de nauwkeurigheid om tot niet arbitraire richtlijnen te komen.

Om hier een beter inzicht te krijgen kan men de tellers $[ur]_B$ en de noemers $[ur]_C$ tegen elkaar uitzetten, zoals in fig. 4 is gedaan (terwille van de duidelijkheid met verschillende schalen op de horizontale en verticale as).

De punten die in deze figuur onder de met I aangegeven stippellijn liggen beantwoorden dan aan de hierboven geëcarteerde waarnemingen. Daarbij werd alleen rekening gehouden met de grootheden $[ur]_C$. We zagen evenwel dat ook kleine waarden van $[ur]_B$ tot vermoedelijk onbetrouwbare uitkomsten leiden. In het bijzonder was er reden een quotiënt q_{ur} te wantrouwen indien zowel de teller als de noemer klein zijn, wat voor die punten van fig. 4 geldt die dichtbij de oorsprong liggen. Een hierop gebaseerde ecartering doet fig. 5 ontstaan.

Ditmaal beantwoorden de weggelaten waarnemingen aan die punten in fig. 4, die liggen onder de (alweer tamelijk willekeurige) stippellijn II.

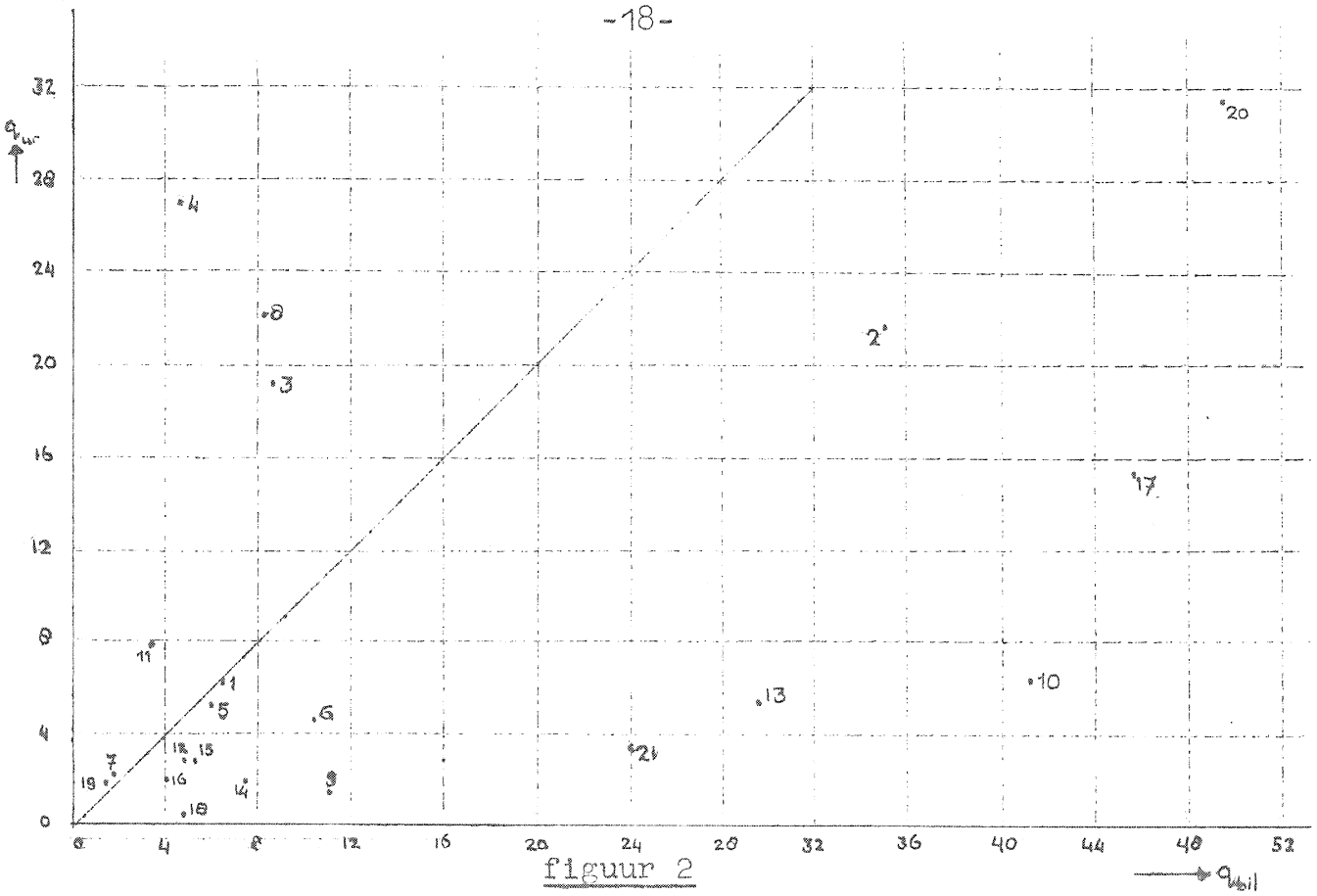
Voor het merendeel zijn de overgebleven punten in fig. 5 dezelfde als in fig. 3. Alle op één na liggen weer onder de bissectrice. Ook nu is de aanwijzing voor een systematisch verschil aanzienlijk duidelijker dan in fig. 2, zij het zwakker dan in fig. 3, vanwege het ene punt ver boven de bissectrice.

Een derde methode komt tegemoet aan het bezwaar dat in fig.3 vrij opzettelijk hooggelegen punten waren geëlimineerd. Men bevordert het tegendeel door juist die waarnemingen uit te zoeken met de grootste tellers $[ur]_B$. Evenals bij fig. 3 zullen ook ditmaal i.h.a. betrouwbaarder punten uit fig. 2 overblijven. Kiest men weer 8 punten, nu de 8 met de grootste tellers $[ur]_B$, dan zijn dit (min of meer toevallig) precies dezelfde als in fig. 5 aangegeven, zodat ook nu weer 7 van de 8 punten onder de bissectrice liggen. Met name liggen de punten, beantwoordend aan de 3 hoogste waarden van de teller, eronder. Onze poging bij voorkeur boven de bissectrice over te houden heeft dus een averechts effect!

Al deze argumenten wijzen dus in dezelfde richting: dat q_{ur} systematisch kleiner is dan q_{bil}, dus dat er afbraak of resorptie van urobilinogenen plaats vindt in de galblaas.

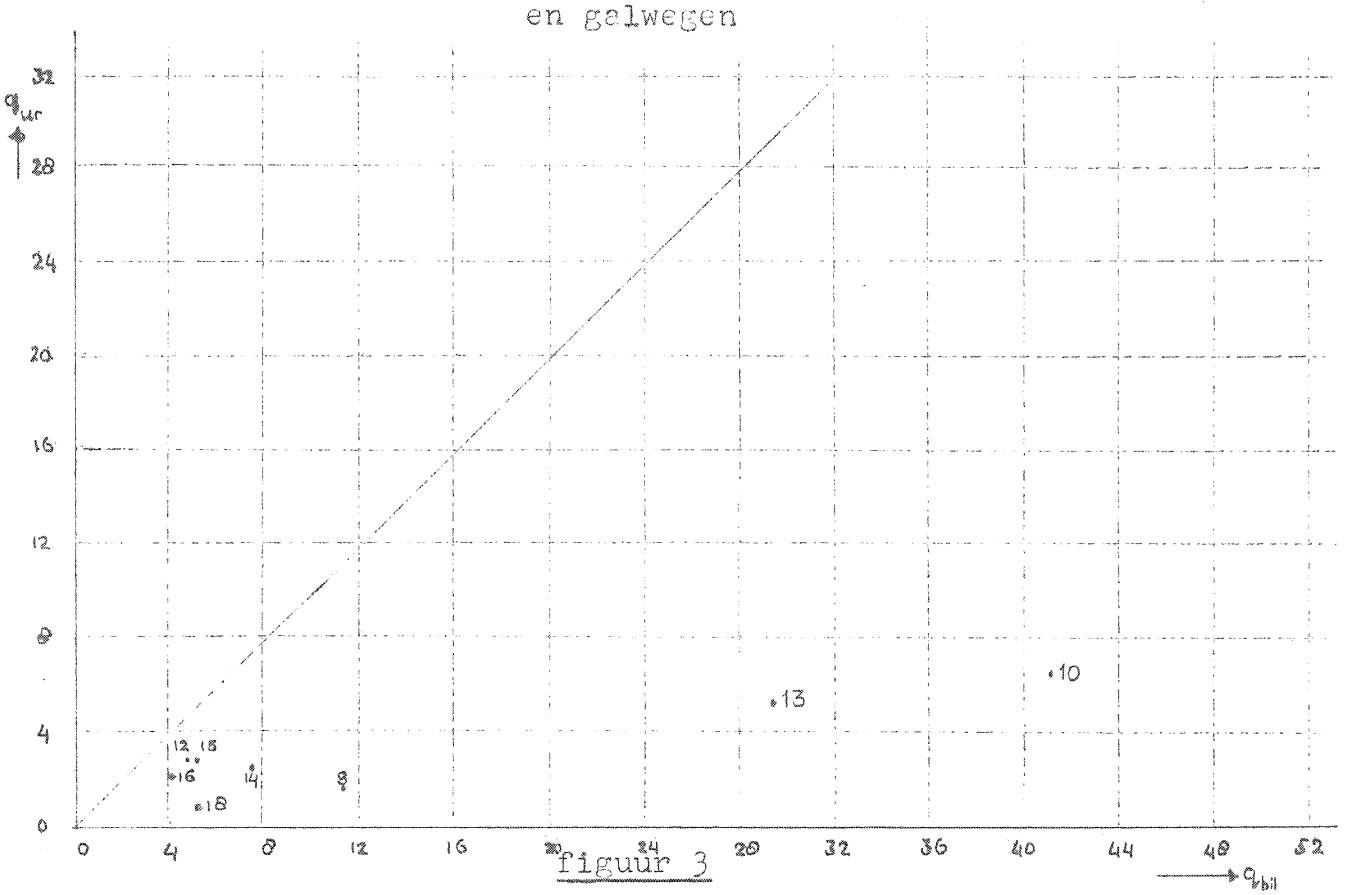
Het speculatieve karakter van deze beschouwingen maakt het

onmogelijk dit als een "bewijs" op te vatten. Men kan slechts de tweeledige conclusie trekken dat mogelijk wel een duidelijke aanwijzing zou zijn verkregen als meer aandacht aan de nauwkeurigheid was geschonken, en dat een verder onderzoek, - indien deze kwestie voldoende belangrijk wordt geacht -, aanbevelenswaardig is.



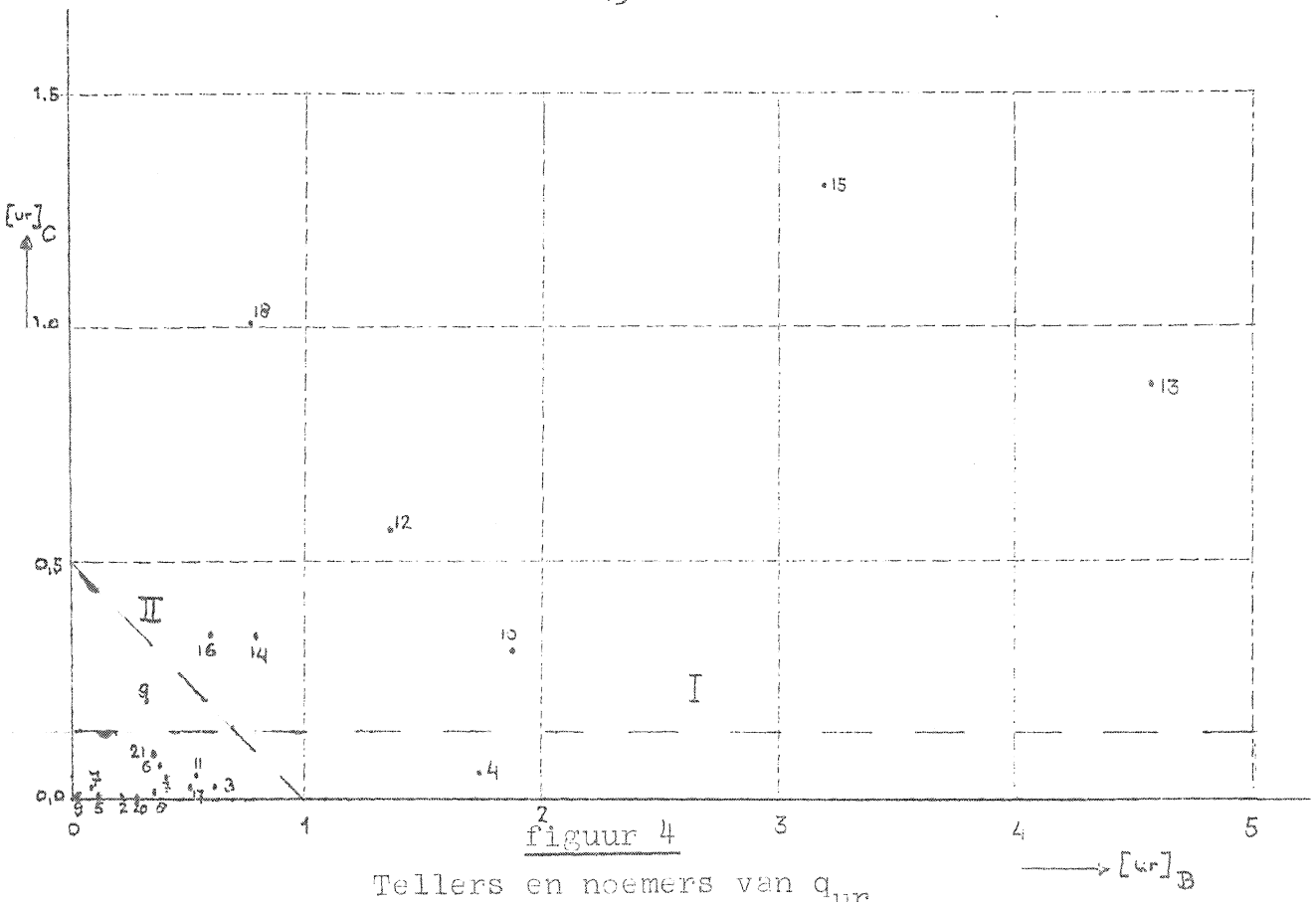
figuur 2

Concentratiequotiënten van 21 patiënten met vrijwel normale lever en galwegen



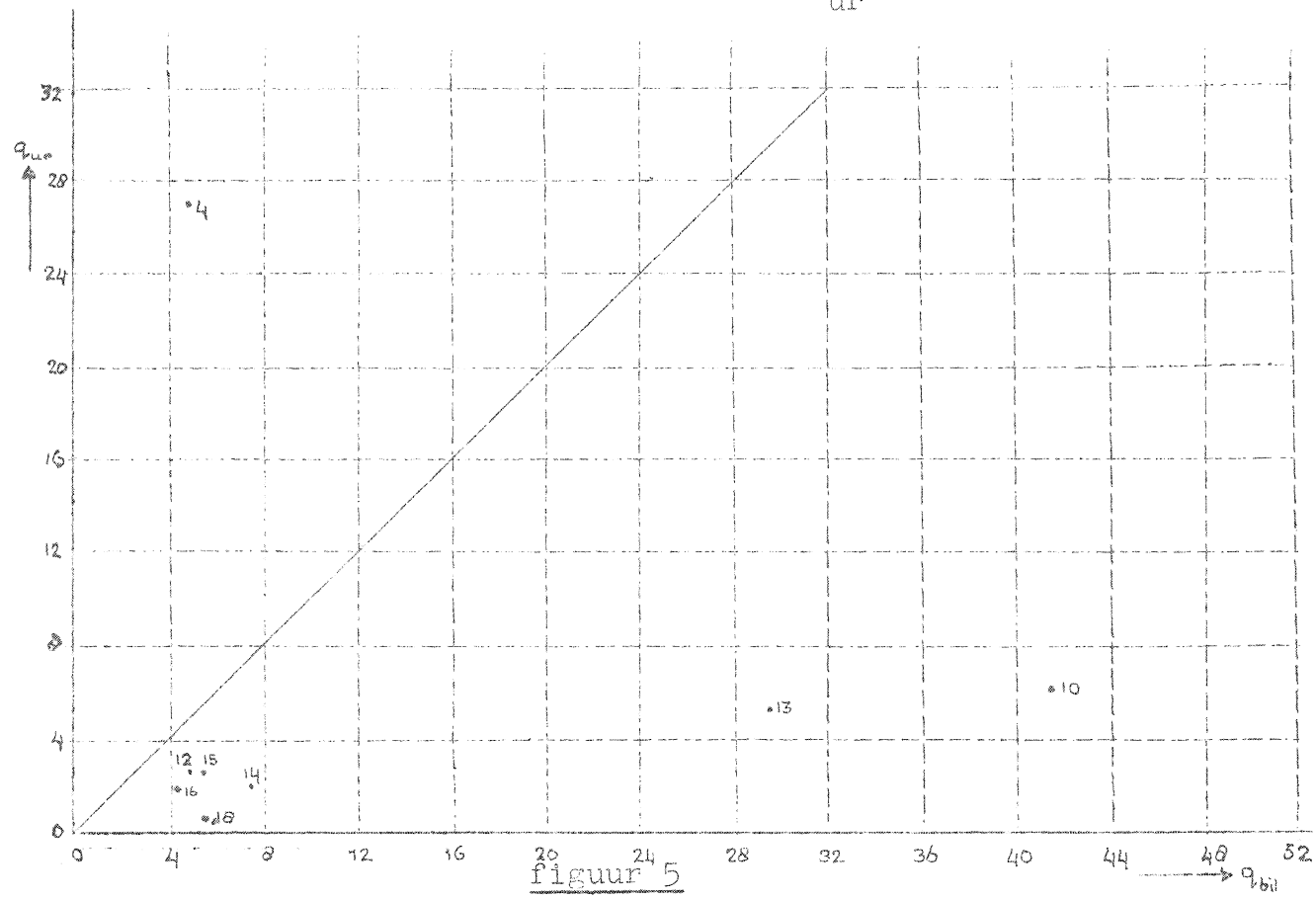
figuur 3

Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_C$



figuur 4

Tellers en noemers van q_{ur}



figuur 5

Concentratiequotiënten van 8 patiënten met de grootste waarden van $[ur]_B + 2[ur]_C$, resp. van $[ur]_B$